

(4)



PATENT APPLICATION

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In application of

Express Mail No. ET161062616US

Richard James GILBERT et al.

Appln. No.: 09/992,440

Group Art Unit: 2171

Confirmation No.: 6119

Filed: November 16, 2001

For: METHOD FOR GENERATING A DATABASE OF MOLECULAR FRAGMENTS

SUBMISSION OF PRIORITY DOCUMENT

Commissioner for Patents
Washington, D.C. 20231

Sir:

Submitted herewith is a certified copy of the priority document on which a claim to priority was made under 35 U.S.C. § 119. The Examiner is respectfully requested to acknowledge receipt of said priority document.

Respectfully submitted,

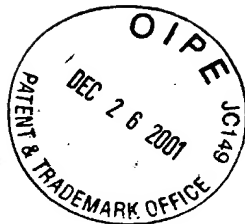
Frank L. Bernstein
Registration No. 31,484

SUGHRUE MION, PLLC
1010 El Camino Real, Suite 360
Menlo Park, CA 94025
Tel: (650) 325-5800
Fax: (650) 325-6606

Enclosures: Great Britain 0028157.6

Date: December 26, 2001

THIS PAGE BLANK (USPTO)



Gilbert et al
US81 09/992 480
filed 11/16/2001
1 of 1



INVESTOR IN PEOPLE

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed 
Dated 28 NOV 2001





The
Patent
Office

1/77

I allow this unstamped form... to
be substituted for the original stamped
document which has been lost in the
circumstances as set out in the Statutory
Declaration of... office

The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

For the Assistant Comptroller.

1. Your reference

PJF01041GB

0028157.6

2. Patent application number

(The Patent Office will fill in this part)

3. Full name, address and postcode of the or of each applicant (underline all surnames)

Amedis Pharmaceuticals Ltd
12 St. James's Square
London
SW1Y 4RB

Patents ADP number (if you know it)

If the applicant is a corporate body, give the country/state of its incorporation

4. Title of the invention

method for predicting a biological target
characteristic of a molecule

5. Name of your agent (if you have one)

"Address for service" in the United Kingdom to which all correspondence should be sent (including the postcode)

Gill Jennings & Every
Broadgate House
7 Eldon Street
London
EC2M 7LH

Patents ADP number (if you know it)

6. If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and (if you know it) the or each application number

Country

Priority application number
(if you know it)

Date of filing
(day / month / year)

7. If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application

Number of earlier application

Date of filing
(day / month / year)

8. Is a statement of inventorship and of right to grant of a patent required in support of this request? (Answer 'Yes' if:

- a) any applicant named in part 3 is not an inventor, or
 - b) there is an inventor who is not named as an applicant, or
 - c) any named applicant is a corporate body.
- See note (d))

Yes.

Patents Form 1/77

9. Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document

Continuation sheets of this form

Description 31

Claim(s) 7

Abstract

Drawing(s) 5

10. If you are also filing any of the following, state how many against each item.

Priority documents

Translations of priority documents

Statement of inventorship and right to grant of a patent (Patents Form 7/77)

Request for preliminary examination and search (Patents Form 9/77)

Request for substantive examination (Patents Form 10/77)

Any other documents (please specify)

11. I/We request the grant of a patent on the basis of this application.

Signature

Date

12. Name and daytime telephone number of person to contact in the United Kingdom

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

- If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.*
- Write your answers in capital letters using black ink or you may type them.*
- If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.*
- If you have answered 'Yes' Patents Form 7/77 will need to be filed.*
- Once you have filled in the form you must remember to sign and date it.*
- For details of the fee and ways to pay please contact the Patent Office.*

METHOD FOR PREDICTING A BIOLOGICAL
TARGET CHARACTERISTIC OF A MOLECULE

Field of the Invention

5 The present invention relates to a method of predicting biological target characteristics of molecules.

Background to the Invention

10 The effect of new and existing molecules upon biological systems is of great importance, in particular with respect to producing new drugs for the treatment of disease.

15 The conventional approach to the development of new drugs has been the synthesis of large numbers of new molecules, followed by the selective testing of the most promising candidates in a series of experimental stages. Such experiments attempt to quantify a particular biological activity of interest for each molecule as this represents the effectiveness of the molecule for its intended purpose. However, experimentation is costly and time consuming and therefore it is particularly desirable to replace experimentation to an extent with appropriate modelling. One method of doing this is to model Quantitative Structure-Activity Relationships (QSAR).

25 The aim of QSAR modelling is to predict a biological activity of interest for previously untested molecules. The predictions are made on the basis of "physico-chemical descriptors". These are fundamental physical or chemical properties of the molecules which may be readily determined or computed.

30 Data according to Selwood *et al.* (*J. Med. Chem.* 1990, 33, 136-142) provides a convenient benchmark for assessing the predictive ability of new models. The Selwood data provides biological activity data for 31 antifilarial antimycin analogue molecules along with data for 53 corresponding physico-chemical descriptors for each molecule. The biological activity used is $-\log(EC_{50})$, that

is the concentration of the molecule required to reduce the release of adenine by 50%.

So and Karplus (*J. Med. Chem.* 1996. 39. 1521-1530) have produced a model to predict the biological activity values using three of these physico-chemical descriptors. This was achieved using genetic algorithms to select the three descriptors. A simple neural network structure (having 3 input nodes, 3 hidden nodes and an output) was then used to predict the biological activity values of the Selwood data. Although the method reproduced the results of earlier findings by others in the field, problems are encountered in the generalisation of this method to predict more complex biological properties such as oral bioavailability, organismal disposition, which are likely to be the result of the interaction of the molecule with many others in the body. There is a need to produce a more versatile method which is capable of generalisation to such properties.

Summary of the Invention

In accordance with a first aspect of the present invention we provide a method of predicting a target biological characteristic of a target molecule, the method comprising:

a) providing a data set for a number of known molecules, the data corresponding to a target biological characteristic and a number of molecular characteristics, wherein the molecular characteristics comprise at least one structural descriptor of the known molecules;

b) selecting a number of the molecular characteristics from the data set;

c) determining a relationship between the selected characteristics and the target biological characteristic;

d) obtaining data corresponding to the selected molecular characteristics for the target molecule; and,

e) predicting the biological target characteristic data for the target molecule based upon the obtained

molecular characteristics data and the determined relationship.

We have realised that a new kind of molecular characteristic can be used to give improved predictions of the biological target characteristic of interest. Known methods utilise physico-chemical data relating to the molecules. Examples of these include partial atomic charges, electro/nucleophilic superdelocalizabilities, dipole moments, Van der Waals volume, surface area, molecular weight, melting point and principal moments of inertia. Unlike these physico-chemical properties, the present invention considers the structure of the molecules using structural descriptors. These are provided in the molecular characteristics data set for selection during modelling, although the data set may also contain more conventional physico-chemical data.

The structural descriptors preferably represent fractional parts of molecules and are not limited to predefined chemical groups. For example one particular descriptor may be defined by two carbon atoms of a benzene ring, bonded to one carbon and one hydrogen atom of a methyl group. The number of structural descriptors that may be defined is therefore very large although some rationalising of these can be performed for example by the eliminating those that occur infrequently.

Structural descriptors are also helpful in understanding the relationships between molecules and the biological target properties that they influence. As the descriptors represent physical fragments of known molecules, if the presence of a particular structural descriptor is found to be important in predicting a particular biological target characteristic, then the atoms within the descriptor's structure along with their bonding and chemical environment can be analysed by chemists to provide information about which parts of the known molecules are active with respect to this characteristic.

The use of structural descriptors also allows the modelling of many different types of biological target characteristic which may include a QSAR activity. Preferably the biological target characteristic will represent a physiological property such as the molecule's absorption, disposition, metabolism or excretion by an organism which is the result of the interaction of the molecule with a number of molecular entities within the organism. Typically the number of such entities will be greater than three. Examples of biological target characteristics include oral bioavailability, plasma protein binding (%), urinary excretion of the unchanged drug, clearance, serum half-life and volume of distribution.

The selection of the particular molecular characteristics and the determining of a relationship between them and the biological target property may be achieved by any known method. Preferably however this will be achieved using predictive numerical modelling techniques and examples include artificial neural networks, rule induction, partial least squares analysis, principal component analysis and evolutionary methods such as genetic programming.

These may involve the steps of selecting the molecular characteristics and determining a relationship between them and the biological target property being performed repeatedly. An evaluation of the predictive power of the determined relationship may be then performed for use in a subsequent selection step. An appropriate fitness parameter may be used for this purpose.

Typically, the selected molecular characteristics will include at least one of the structural descriptors.

In general only a relatively small number of the molecular characteristics may be useful in the prediction of the biological target property. Preferably the molecular characteristics will be selected using an explicit modelling system such as an evolutionary

algorithm. Any known evolutionary algorithm method may be used including evolutionary programming or a genetic function approximation method.

5 The determining of a relationship between the molecular and biological target characteristics will be preferably achieved using artificial neural networks due to their modelling flexibility. Advantageously the evolutionary algorithm may be used in the selection of the molecular characteristics and at least one artificial
10 neural network parameter for use in accordance with the artificial neural network. This parameter may be used for example to control the number of hidden nodes in the artificial neural network. Alternatively or additionally, the parameter may correspond to a regularisation parameter
15 for the artificial neural network.

To improve the predictions, the method may be performed a number of times using different starting conditions. The method may then further comprise selecting a number of groups of molecular characteristics having
20 corresponding determined relationships and combining the predictions of each selected group and corresponding relationship to produce a prediction for the biological target characteristic of the target molecule.

Typically the method will be implemented using a
25 ~~system such as a computer under the control of a computer~~ program comprising program code means adapted to perform the method. The computer program may be typically embodied on a computer readable medium.

In accordance with a second aspect of the present
30 invention, we provide system for determining a molecular structural descriptor from molecular coding data contained within a data set, the system comprising:

a data store for storing a data set for a number of known molecules, the data corresponding to a target
35 biological characteristic and a number of molecular characteristics, wherein the molecular characteristics

comprise at least one structural descriptor of the known molecules;

5 a program store for storing the computer program, the computer program comprising instructions for performing the methods relating to the first aspect of the invention; and
a processor for executing the computer program contained within the program store.

10 One known method of predicting QSAR activities according to So and Karplus involved the selection of three physico-chemical descriptors using a genetic algorithm. Artificial neural networks were used to evaluate the predictive power of the selected descriptors. However, with this method the artificial neural network was rigidly predefined based upon the data set. In addition a network
15 smoothing regularisation parameter was also assumed from an analysis of the data.

In accordance with a third aspect of the present invention, we provide a method of predicting a biological target property of a target molecule, the method comprising
20 the steps of:-

a) providing a data set for a number of known molecules, the data corresponding to a target biological characteristic and a number of molecular characteristics;

~~25 molecular characteristics of the data set and the~~
b) determining a relationship between a number of the biological target property of the data set by repeatedly performing the steps of:-

i) selecting a number of groups of molecular characteristics;

30 ii) selecting at least one artificial neural network parameter for each group;

iii) for each group, using an artificial neural network in accordance with the selected artificial neural network parameter to determine a relationship between the
35 molecular characteristics data and the biological target property data of the data set;

iv) assessing the performance of the respective artificial neural network for each group;

v) repeatedly performing steps (i) to (iv) using the assessed performance in the selection of molecular characteristics and artificial neural network parameters in
5 subsequent steps (i) and (ii); and

c) predicting the biological target property for the target molecule using the determined relationship and the respective molecular characteristics of the target
10 molecule.

This provides advantage over prior methods in that the artificial neural network (ANN) has an in-built flexibility. The method is not restricted to the selection of a predetermined number of molecular characteristics for
15 each group and accordingly there is no predefined assumption as to the specific form that any determined relationship will take. Unlike in previous methods the number of molecular characteristics used in the final relationship and the structure of the corresponding
20 artificial neural network are selected iteratively.

Each ANN is allowed to evolve alongside the evolution of selected groups of molecular characteristics. This is due to the selection of at least one artificial neural network parameter for each group. In this way more
25 ~~accurate relationships may be determined as the method~~ allows a more comprehensive search of multi-peaked and rugged fitness landscapes to be performed.

The target biological property according to all aspects of the invention may comprise an activity such as
30 found in QSAR studies or a physiological property.

The number of molecular characteristics in the data set may be greater than the number of molecules in the data set. The ability to consider large numbers of molecular characteristics is advantageous particularly when limited
35 knowledge exists as to the magnitude of the biological effect in question.

The molecular characteristics data may include fundamental physico-chemical data and structural descriptors for the molecules. In addition *in vivo* and/or *in vitro* data may be used.

5 Preferably the selected ANN parameter will be the number of hidden nodes of the ANN and typically a further regularisation parameter will also be selected in each case.

10 Preferably the iterative selection of groups of molecular characteristics and ANN parameters will be performed using a known evolutionary algorithm method. During the iteration process, the predictive performance of each selected group (including the corresponding ANN parameters) will be preferably assessed using a fitness
15 parameter.

Typically a fitness ranking may be used to determine which groups of molecular characteristics will be used in the formation of further groups, those groups exhibiting a higher fitness having a corresponding higher likelihood of
20 selection.

This provides particular advantage in that a choice of the regularisation parameter can be made based upon the fitness parameter for the groups. Contrary to previous methods, a detailed analysis of an appropriate value for
25 the regularisation parameter is therefore not required as this is inherently chosen over successive generations.

Preferably following selection, each group will be subject to a possible "mutation" according to probability distributions.

30 Typically a mutation will involve the random replacement of one or more molecular characteristics of a group with a molecular characteristic from the data set, selected at random. Similarly one or more molecular characteristics may be added or removed at random from a
35 group. Mutation of the ANN network parameters may also be performed, again according to a distribution.

The predictions of a number of selected groups having corresponding relationships may be combined as in the case of the first aspect in order to produce a prediction for a target molecule.

5 Similarly to the first aspect of the invention, this aspect will be preferably implemented using a computer system under the control of a computer program which may be stored on an appropriate computer readable medium.

10 In accordance with a fourth aspect of the present invention we provide a system for predicting a biological target property of a target molecule, the system comprising:

15 a data store for storing a data set for a number of known molecules, the data corresponding to a target biological characteristic and a number of molecular characteristics;

 a program store for storing a computer program, the computer program comprising instructions for performing the method described relating to the third aspect; and

20 a processor for executing the computer program stored in the program store.

 In accordance with a fifth aspect of the present invention, we provide a method of determining a molecular structural descriptor from a molecular coding data set
25 ~~contained within a store, the method comprising:~~

 a) accessing the store to select first molecular coding data from the molecular coding data set, the first molecular coding data describing a first molecular structure;

30 b) accessing the store to select second molecular coding data from the molecular coding data set, the second molecular coding data describing a second molecular structure;

35 c) processing the selected first and second molecular codings data to determine common structural coding data representing a common molecular structure between the first and second molecular structures; and

d) storing the determined common structural coding data in the store, the common structural coding data representing the molecular structural descriptor.

5 The molecular coding data may comprise data describing the chemical moieties within the first and second molecules and the structural descriptors, along with data describing the bonding and atom types of these moieties.

10 In general the method will further comprise repeatedly performing steps (a) to (d) above, upon the coding data in the store, such that common structural coding data is determined for all of the molecular coding data and common structural coding data in the store related to the molecular coding data set.

15 The step of processing the first and second molecular codings data will typically comprise the steps of:

i) converting the first and second molecular codings data to first and second coloured graphs according to graph theory;

20 ii) determining a docking graph from the first and second graphs;

iii) identifying at least one clique within the docking graph; and

iv) converting each clique identified into common structural coding data.

25 Although all identified common structural coding data may be stored, typically the data will only be stored if it represents a unique molecular structure with respect to the corresponding molecular structures of the coding data already contained within the store.

30 To further reduce the amount of data used, the common structural coding data may be ranked according to the frequency with which it is identified within the molecular coding data in the store. The common structural coding data which occurs less frequently than a predetermined frequency threshold may then be discarded.

35 Preferably, the determined common structural coding data will be common substructural coding data, relating to

common molecular substructures identified from the data within the data set.

5 The store may be of any known configuration and may be located remotely from the means provided to perform the processing of the coding data. In addition the store itself may be divided between a number of stores, possibly in separate locations and linked by appropriate communication means.

10 Preferably, the method will be used in accordance with the method and apparatus of the first and second aspects of the present invention.

In accordance with a sixth aspect of the present invention we provide a system for determining a molecular structural descriptor from molecular coding data contained within a data set, the system comprising:

15 a data store containing a molecular coding data set;
a program store for storing the computer program, the computer program comprising instructions for performing the method described relating to the fifth aspect; and
20 a processor for executing the computer program contained within the program store.

Any of the systems described above may further comprise an input means to enable a user to control the system and enter data for use by the computer program.
25 ~~Typically such data will comprise molecular characteristics~~
data relating to new molecules. In each case the system may be provided with communication means in order to allow the user to control and access the system from a remote location, for example using the Internet.

30

Brief Description of the Drawings

An example of a method for predicting a biological target property will now be described with reference to the accompanying drawings, in which:-

35 Figure 1 is a flow diagram overview of the example;

Figure 2 is a flow diagram of the genetic algorithm method of the example;

Figure 3 is a schematic illustration of the production of new genotypes according to the example;

Figure 4 is a flow diagram of the artificial neural network method of the example; and,

5 Figure 5 is an illustration of apparatus according to the example.

Detailed Description

Referring to Figure 1, a number of known molecules are
10 selected at step 1 for use in a prediction data set. The data set can be thought of as a spreadsheet comprising a number of rows, each row representing a specific molecule and a number of corresponding columns containing biological target characteristics and molecular characteristics data
15 relating to each particular molecule. The purpose of the data set is to provide a knowledge base for the modelling of a biological target characteristic, that is the characteristic to be predicted for molecules not contained within the data set.

20 For each molecule in the data set, data (such as a measured value) for the biological target characteristic is known, as is data describing the molecular structure and other related physico-chemical data. The purpose of the model is to establish a relationship to predict the
25 ~~biological target characteristic~~ (which may include a conventional QSAR activity) based upon selected molecular structure and/or physico-chemical data.

A number of biological target properties may be known for each molecule and each of these may be predicted
30 separately by individual models.

In this example the first column represents the known target biological characteristic of interest. This may be thought of as a "high level" biological property, that is a property relating for example to the interaction of the
35 molecule with a biological system.

An example of a high level biological property is "urinary excretion in humans" which is a measure of the

tendency of a particular molecule be excreted in the urine unchanged, rather than being broken down or eliminated by another route.

The majority of the remaining columns represent a number of the molecular characteristics of the molecule. These include structural descriptors and more conventional physico-chemical properties such as those used in the Selwood data set. These molecular characteristics may be collectively considered as "low level features".

The structural descriptors are particularly important in that they represent a new approach in the modelling of biological characteristics. A large number of these may be defined, generally comprising structural groups of molecules and in particular these may include fragments of molecules which would not be normally considered as single structures.

A large number of structural descriptors can be defined in the following way. Firstly a chemical moiety is selected which may comprise atoms such as C, N, O, Si, P, S, H, F or halogens; or bonded groups such as SO₂, CO, CN, NO₂ or a benzene ring. Each of these has a specific valency and may/may not support double bonding. Secondly, in accordance with possible double bonding and valency, the combinations of moieties from the same list to which the first could be bonded are calculated. Each of the combinations of the first and second moieties represents a structural descriptor (subject to certain chemical bonding rules). These descriptors may then be used in the analysis of molecules. The list of moieties may be expanded to further atomic elements or chemical groups, as can the number of moieties considered within a descriptor. However, a large number of these descriptors may not be found in a particular set of molecules of interest.

Using this approach, to analyse a new molecule with a known structure, the structure is searched for any moieties corresponding to the first set. Those located are then analysed in terms of the moieties to which they are bonded

and the types of bonds present. These are then identified with respect to the descriptors defined above. The frequency of occurrence of these descriptors is then computed as is the frequency of individual bond types between them. As an additional step, the frequency of three-atom combinations of the form A-X-B is also computed if the atom X has a valency of three or more. In this manner the new molecule can be "coded" in terms of descriptors.

An alternative to the above is to search for structural descriptors within known molecular structures, for example those molecules within the data set. This can be achieved by performing the following steps, which involve the methods of graph theory:

1) Compile a structure library of the molecules of interest, the library including the molecules described in the data set which will be used for predictions. Typically the structure of molecules is stored with reference to the three dimensional coordinates of the atoms along with the identities of the other atoms within the structure to which they are bonded.

2) Select any two molecular structures from the library.

3) Convert the two molecular structures into corresponding "coloured" graphs (connectivity tables) with bonds as nodes and atoms as edges. For each atom in each

molecular structure, evaluate the number of bonds to that atom, then deduce the type of bonding involved (for example a double bond), analyse the atomic number of the actual atom, and finally evaluate the chemical context of the atom. The chemical context provides information upon the environment surrounding the atom, particularly in terms of the atoms/groups to which it is bonded. For example the carbon atom at the centre of a ethylene group would have a different chemical context to a carbon atom in carbonyl group, despite both atoms having two single and one double bond. The chemical context is deduced in each case by considering the particular atom involved, the number of

bonds and the atoms to which it is bonded, the chemical context then being deduced from a suitable look-up table.

4) "Reduce" the graphs for each molecule in order to produce in each case the minimum graph required to describe the respective molecule fully.

5) Create a "docking" graph by analysing the two graphs. The docking graph represents the structures that are common to the two graphs. This involves picking a bond in the first molecule and one in the second, comparing them and if they connect similar atoms, then looking at the other bonds of those atoms, comparing them and so on such that the structures of the two molecules are searched fully and the structures common to both molecules are maximised.

6) Scan the docking graph to identify "cliques" which are unconnected regions of substructure common to both molecules. A suitable method for achieving this is described in "Algorithm 457: Finding all cliques of an undirected graph", C. Bron and J. Kerbosh, Communications of the ACM, volume 16, number 9, 1973, pages 575 to 577.

7) Store each clique located.

8) Convert the cliques back to molecular structure fragments by copying the respective structural details from the original molecules.

9) Calculate the molecular mass of each new fragment (clique) and compare it with those already found. If the fragment has a unique mass, skip to step 12. In some cases, particularly with larger fragments, there may be a number of possible structures which have the same mass and it is desirable to identify each of these structures whilst eliminating identical structures. Therefore, if two have the same mass, proceed to step 10.

10) Repeat steps 3-8 using the two fragments with common molecular masses and treating them as "molecules".

11) If the mass of the maximum common substructure between the two fragments is the same as the mass of the fragments, then they are exact duplicates, and one can be discarded.

12) Add the (now unique) fragments to the original structure library. Therefore although initially the structure library will contain only the original molecules, as the search for fragments proceeds, the number of
5 "molecules" in the library also increases.

13) Repeat steps 9-12 for all new fragments found.

14) Repeat steps 2-13 using all pairwise combinations of the original structures plus newly-found fragments until no new fragments are found.

10 Using this approach, all fragments common to at least two molecules are found. Many fragments can be eliminated by discarding infrequently occurring ones (i.e. ones that are found in only a few molecules) as this reduces the computational demands of the model.

15 It is these fragments that are used within the prediction data set as structural descriptors.

The size of these structures can range from a pair of bonded atoms up to the size of the maximum common substructure. The advantage of this graph theory approach
20 is that structural descriptors are only found that exist within the molecules of the data set. In practice both methods described above may be used in combination.

Examples of the more conventional molecular characteristics which may be used include the molecular
25 ~~mass, the molecular volume, surface area, moments of~~
inertia, melting point and electronic superdelocalizabilities. Much of these conventional data may be obtained by known computation techniques or from reported experiments within relevant literature.

30 Many of these molecular characteristics will be in the form of integers. For example the number of times that a particular structural descriptor is found within a molecule may be represented with an integer in an appropriate column. Other molecular characteristics may represent
35 certain choices from a predetermined range, each possible choice being represented by an integer. Some of the molecular characteristics, particularly the fragment counts

and typically the biological target characteristic, will be a real value. A typical data set may comprise approximately 200 molecules (rows) and around 1000 molecular characteristics (columns).

5 Returning now to Figure 1, the data representing the molecular characteristics are entered into the spreadsheet at step 2.

10 Using the method of the present example it is not necessary for the number of rows to exceed that of the columns. However, using alternative modelling techniques to those of the present example, there may be too many variables with respect to the amount of data provided if the number of columns exceeds that of the rows.

15 In this example it is assumed that only a relatively small number of the characteristics will be important in the prediction of the biological target characteristic. The problem will therefore be soluble if the number of molecular characteristics selected is equal to or less than the number of molecules in the data set.

20 There are however two major problems to be addressed with this approach. Firstly, which particular molecular characteristics are important is not known, and secondly the manner in which these characteristics interrelate is also not known.

25 ~~The present example provides a method of selecting the relevant molecular characteristics and determining a relationship between these and the biological target characteristic.~~

30 In step 3 of Figure 1 the spreadsheet is prepared with one row for each molecule, a column for the target biological characteristic data and successive columns for the data relating to each respective molecular characteristic.

35 The data set is then pre-processed at step 4. This involves the removal of redundant columns in which there are no data and a cross-correlation check in which columns are compared and one is deleted if a greater than 90

percent correlation is found between them. Further pre-processing may include the normalisation of some of the columns and replacing the data in others with logarithmic values.

5 The model itself combines the use of an evolutionary algorithm in the form of a genetic algorithm (GA), with trained artificial neural networks (ANN), in order to assess which molecular characteristics can be successfully used to predict the biological target characteristic.

10 A GA effectively provides the flexibility in the model to allow the choice of varying numbers of molecular characteristics, whereas the neural network attempts to find the best fit relationship (which generalises to unseen data) between those chosen and the target characteristic.

15 However, the genetic algorithm not only selects the molecular characteristics and uses the results of the neural network analysis to guide its further selection, it also allows some modification of the neural network in order to add greater flexibility to the model.

20 In Figure 1 at step 5 the molecules (rows) in the data set are divided up into groups of n rows (for example n being 5). This is for the ANN cross-validation purposes to be described later.

25 The artificial neural networks (ANN) are then initialised at step 6. This includes a selection of the type of ANN to be used along with a selection of the noise function defining the assumed noise in the biological target characteristic data (for example Gaussian or Beta). Typically, a non-linear ANN model is chosen with a variable architecture and assumed Gaussian noise distribution. A Beta distribution may be used for the noise at this stage although this is computationally more demanding.

30 The genetic algorithm (GA) and the artificial neural network (ANN) are then presented with the data set at step 35 7. Figures 2 and 4 describe the operation of the GA and ANN respectively.

Th Genetic Algorithm

Referring to Figure 2, the genetic algorithm of the present example has three main functions as indicated. The first is to select a number of groups of molecular characteristics from the data set and, in accordance with the neural network model, to produce improved selections of groups with correspondingly improved predictions of the target biological characteristic.

A second and third function of the GA is to choose an appropriate number of hidden nodes for the ANN and an appropriate regularisation coefficient for ANN smoothing such as a "weight decay" parameter.

A selection of these ANN parameters is made in association with each group. Their selection by the GA provides particular advantage with respect to known methods in that it gives the ANN enhanced flexibility allowing more accurate predictions to be made and addresses the problem of the ANN overfitting to the training data through the evolutionary selection of an appropriate "weight decay" parameter. This is described in more detail later.

The GA defines a population P of "genotypes" (typically under 100) at step 8, each of which represents a subset of possible molecular characteristics and is associated with a number of hidden nodes for the ANN, and an ANN regularisation parameter. The genotype molecular characteristics are represented by integers, each integer representing a corresponding column in the data set.

In this example 50 genotypes are chosen to form the population P, each member of which is to be evaluated by a respective ANN to give a "fitness". The fitness is a measure of the accuracy with which the biological target characteristic can be predicted upon unseen data using the selected molecular characteristics of the particular genotype.

Initially at step 9, the 50 genotypes in the population P are generated. The creation of each genotype

involves the selection of a number of molecular characteristics from the data set, along with the selection of the number of hidden nodes for the ANN and an ANN regularisation value. The number of molecular characteristics for each genotype of the population will not necessarily be equal, rather being selected at random from an appropriate distribution. The mean of this distribution may be selected iteratively by the model operator depending upon the results obtained. The initial selection of each particular molecular characteristic is also made at random. However, it is ensured that no particular molecular characteristic is selected twice within one genotype.

The two ANN parameters are also chosen from appropriate distributions, for example for one particular data set the number of hidden nodes may be chosen at random from an integer range 1 to 20. Similarly real values may be used for the ANN regularisation parameters.

At step 10, the predictive performance of each genotype within the population P is evaluated using an ANN which returns a "fitness" value. The detailed architecture and operation of the artificial neural networks between steps 11 to 19 is described in more detail later with reference to Figure 4.

Following fitness evaluation, the population P members are then ranked in accordance with their assessed fitness at step 20. Using this ranking, pairs of "parent" genotypes to act as parents for the next generation are then generated. This is achieved in the following way.

A pool of parental genotypes is created from the population P using a stochastic universal sampling (SUS) method due to Baker described in "Reducing bias inefficiency in the selection algorithm" (Proceedings of the 2nd International Conference on Genetic Algorithms, pl4-21, ed. Grefenstette J. J., Lawrence Erlbaum Associates, Hillsdale NJ (USA); 1987) and incorporated herein by

reference. Of course, other selection regimes could be used.

Using this method the highest fitness genotype is represented within the pool by some number $1+N$ copies. In the present example N is 1 so there are two copies. For linear ranking the median fitness type is represented once, and the least fit genotype is represented $1-N$ times (or where $N=1$ not represented at all). Intermediate genotypes are represented with an expected number that is linearly interpolated according to place within the ranking. The expected number typically consists of an integer (including zero as an integer) plus a fractional part, and the SUS method ensures that at a minimum, the integral number is guaranteed to appear within the pool, with an expectation of one extra representative with a probability equal to the fractional part. The probabilities are calculated so as to ensure that the parent pool contains as many genotypes as were in the population P .

The parents are then selected from the pool at random in pairs until all the genotypes have been selected. When a genotype has been selected it is removed from the pool. The genotypes having a higher fitness are therefore preferentially selected due to the higher probability of their presence in the parental pool.

Alternatively an "elitist" option can be used to ensure that at least one copy of the high scoring member of any generation is passed on and furthermore that it is unchanged by any later "cross-over" or "mutation" to the following generations.

Pairs of offspring genotypes are then generated from the pairs of parent genotypes. There are a number of different methods for producing offspring genotypes which would generally involve the selection of molecular characteristics from each parent genotype. In the present example, a "bag-crossover" method is used as shown in Figure 3. This can be thought of in the following way:-

The first parent genotype 100 is deconstructed and each integer representing a molecular characteristic is hypothetically placed in a bag 102. In the Figure, the characteristics are represented by letters. A similar process is performed for the second parent 101 although any duplication is avoided. This is shown with respect to characteristic "F" which is present in each genotype 100,101 but is only present in the "bag" 102 once. The bag is then metaphorically shaken and two sets of molecular characteristics are copied from the bag at random to form the two offspring genotypes 103,104. A particular molecular characteristic may therefore be present in both offspring genotypes but no duplication within one genotype is permitted.

This method places no significance in the order of the characteristics and has been found to be effective even when the number of the characteristics differs between the two parents.

The number of molecular characteristics selected from the bag for each offspring may not be necessarily the same as one or both parents. As the parent genotypes may differ in length from one another, the length of each offspring genotype is selected at random from a symmetrical distribution with a mean set at the average number of molecular characteristics of the parents.

For each offspring genotype the number of hidden nodes and the regularisation parameter are each copied randomly from either of the parents as illustrated in the Figure.

Two mutation operators may then act upon the offspring genotypes, according to predetermined probabilities. The first is a replacement operator which acts upon the molecular characteristics and allows any molecular characteristic to be replaced at random by any of the molecular characteristics from the spreadsheet (without duplication), albeit with a small probability of around 1 to 2 percent. This is shown for genotype 103 where the

molecular characteristic "B" is replaced by characteristic "T" from the data set to create the mutated genotype 105.

5 A genetic operator then allows any molecular characteristic to be added or to be deleted at random from the genotype, again with some small probability. Hence the number of molecular characteristics may change through successive generations as influenced by the selection process. This is shown for genotype 104 which becomes genetically modified to genotype 106 due to the addition of
10 the molecular characteristic "R" from the data set.

The two ANN parameters are also subject to mutation operators. This is achieved for both ANN parameters by adding or subtracting at random a value selected from appropriate distributions (typically normal and log normal
15 for hidden nodes and regularisation respectively). For example the number of hidden nodes 50 from parent genotype 100 is mutated to a new number indicated by 54 in genotype 105.

Returning to Figure 2, the selection of new offspring
20 genotypes is shown at step 21, whereas the mutation operators are shown at step 22. Steps 10 to 22 form a loop in which new generations of genotypes are repeatedly created, assessed by the ANN, ranked and selected to produce further generations.

25 The number of generations of genotypes is generally predetermined by the operator of the model, typically a few hundred may be used although the value is very specific to the problem being addressed. The actual number is typically chosen iteratively by the model operator based
30 upon a consideration of the computational time required with respect to the resultant improvement in the model.

After a number of generations, the fitness selection method employed produces a population having the same number of genotypes as the initial population but with
35 typically a few high scoring genotypes (step 23). It is expected that the final population will contain some low fitness genotypes as otherwise this would indicate that the

level of mutation between generations was too low thus not giving satisfactory flexibility in the search for high fitness solutions.

At step 24, the one or more high scoring genotypes
5 defining good ANN models in terms of selected features, hidden nodes and regularization coefficients, are output for later use.

The Artificial Neural Network (ANN)

10 Figure 4 shows a flow diagram of the neural network operation in more detail.

A number of different network architectures and training procedures are known for finding relationships between input node data values and the target values of the
15 output node or nodes. The present example uses a multilayer perceptron ANN architecture with weight decay regularisation.

Although a different ANN is constructed for each genotype, they all conform to the structure of a three
20 layer feed forward network with an input layer, a hidden layer and an output layer, each layer having corresponding nodes. The nodes effectively represent input/output variables for use with functions, the functions defining the connections between the nodes.

25 The input layer comprises a number of input nodes with a node representing each molecular characteristic of the particular genotype. The number of input nodes is equal to the number of molecular characteristics in the genotype plus an additional bias node.

30 The number of nodes within the hidden layer is defined by the respective ANN parameter associated with the particular genotype. The hidden layer may well include more than one node, the greater the number of hidden nodes the greater the potential non-linearity of the mapping. In
35 this case the number of hidden nodes is equal to the integer value parameter associated with the genotype plus a further bias node.

In general, there is often only a single output node representing the predicted mean for Gaussian noise networks. Two output nodes are often used for Beta noise or indeed Gaussian noise with an additional variance
5 output.

The known function of the bias nodes in the input and hidden layers is to allow the weights to be biased. The network weights are coefficients which represent the perceived importance by the network of particular nodes
10 with respect to their counterparts.

The ANN for each genotype is set up at step 11 of Figure 4. At step 12, the weights are set at random from an appropriate distribution. A zero mean unit variance isotropic Gaussian distribution is used in the present
15 example. The variance is scaled with the "fan-in" of either the hidden or output nodes as appropriate. This is performed to reduce the variance of the values between the nodes as typically only a small number of nodes are used.

When the number of input and hidden nodes has been
20 selected along with the regularisation parameter, the ANN is trained and tested upon the data at step 13.

A common problem for ANNs is that of potentially overfitting the training data such that the model does not generalise well to unseen data, which is the ultimate
25 objective. A number of known techniques can be used to balance the accuracy of the fit against the problem of over fitting. A common method of addressing this problem is to perform "cross validation" where part of the data set is put aside and not used to train the ANN. These data are
30 then used as unseen data against which the predictions of the ANN can be assessed and the ANN adjusted accordingly.

In the present example this actually involves the training and testing of 5 individual ANNs using the division of the data performed at step 5 of Figure 1. In
35 each case the ANN is trained upon four of the five groups and tested upon the remaining one. Five ANNs are used such that the groups may be rotated, the result being that each

molecule is used within four networks as part of training data and within one as testing data. The five ANNs are related insofar as they each share the same genotype, inputs, number of hidden nodes and regularisation parameter. However, they differ in that they are trained upon different subsets of the data, resulting in different weights and in addition, they are tested upon unseen data that differs for each ANN.

As shown at step 13, during training the relevant data are fed through the network from the input nodes to the hidden nodes. In the present example, each is passed through a hyperbolic tangent function with a respective bias as is known in the art. The values are then transferred to the output node in a linear manner.

At step 14, for each molecule the predicted output node value is compared with the known target value of the biological characteristic from the data set resulting in a "fitness". This is measured using a correlation coefficient " r^2 " based upon the sum of the squares of the difference between the target output and the predicted output for each molecule. The result is used to adjust the weights and biases so as to reduce the error using a back propagation method at step 15. Typically a "scaled conjugate gradient descent" algorithm is used because of its known efficiency.

The tendency of the ANN to over-fit the data, can be reduced by using various regularisation methods. Here a "weight decay" regularisation method is used. However, the selection of the value of this parameter poses a problem for many ANN methods, small and large values giving over-fitting and under-fitting to the data respectively. Here this is avoided by allowing the GA to choose a suitable value for this parameter in an evolutionary manner along with the number of hidden nodes and the molecular characteristics. This can be done because the ANNs are only tested upon unseen data and therefore the fit to such data is desired to be maximised.

The training is an iterative process and steps 13 to 16 of Figure 4 are performed many times in a loop in order to train the respective ANNs. The number of iteration loops to be performed may be chosen by the operator of the model. This may be based upon the improvement in the fitness as a function of iterations or upon computational time.

At step 17 five trained ANNs are produced for a particular genotype which are then tested upon their respective test data sets at step 18. The cross-validation correlation coefficient on the test data for the five ANNs in combination provides the overall fitness for the genotype and this is ultimately used by the GA in its selection process (step 19).

Returning to Figure 1, the highest fitness genotypes with associated ANN parameters are returned for a particular population at step 7.

To increase the effectiveness of the model, the GA and ANN methods described above are performed on a number of different starting populations, either serially or in parallel such that one or more high fitness models are produced for each population.

The output values of these ANNs represent the mean expected values of Gaussian distributions. However, a Gaussian distribution, even when arranged about a positive mean value is asymptotic in each direction and therefore inherently includes some negative values even for a positive mean. In predicting biological target values, often the target value must fall inherently between upper and lower bounds. Therefore a Beta distribution may be more appropriate where reasonable lower and upper bounds can be set in advance of training.

An optional stage may then be performed at step 26, where each of the high scoring ANNs for each population may be retrained using Beta noise. This is to take account of the bounded distribution of allowed values for the biological target characteristics. However, Beta noise is

computationally more expensive and time consuming but in this case at least only occurs with respect to the high fitness models. The retraining also includes the re-evaluation of the fitness values. Of course Beta noise
5 could be used throughout, given sufficient computational resources.

At step 27 (with or without retraining) one or more high fitness genotype models exist with respect to each starting population. In many cases the fitness values are
10 similar between these models. In order to produce an overall output prediction for the biological target property, a number of these models are then selected for use in a prediction committee. For example the top ten models regardless of their originating population are
15 selected for use in the committee. The committee combines the individual predictions of the committee members to produce an overall output prediction.

To make a prediction, a new molecule X is firstly selected. Appropriate calculations are then performed to
20 provide values for the corresponding molecular characteristics in the data set (although only those used by the committee members are required).

For the molecular characteristics related to structural descriptors and in the case of the method
25 described using graph theory, this is achieved in accordance with the structure library and graph theory method described earlier by performing the following steps:

- a) Select a molecular structure (descriptor) from the library.
- 30 b) Count the number of bonds in the selected structure and store this value in N.
- c) Convert both the selected descriptor structure and the structure of molecule X into coloured graphs.
- d) Reduce the graphs.
- 35 e) Create a docking graph from these reduced graphs.
- f) Scan for cliques within the docking graph.

- g) Identify the largest common structure (clique) between the descriptor structure and the molecule X structure, and set S to the size of this common structure, where S is the number of bonds in the clique.
- 5 h) Add S/N (the proportion of the particular descriptor structure in molecule X) to the total for that particular descriptor structure, as there may be more than one complete/partial occurrence of that descriptor structure within the molecule X.
- 10 i) Set the bond types of the identified common structure in the molecule X to "unbonded" (ie remove the found common structure from the molecule).
- j) Repeat steps (c) to (i) until no more common structures using this descriptor structure are found in molecule X.
- 15 k) Repeat steps (a) to (j) for all descriptor structures in the library.

The descriptor structure totals will then represent the number of times each fragment (or part of a fragment) occurs in molecule X.

20

These values are then entered into each of the genotype models of the committee, each model producing an output value according to its trained ANN.

The output values are then simply averaged and a predicted error may also be calculated in order to produce a final predicted output value.

25

It is possible that some of the high scoring genotypes may have selected the same subsets of molecular characteristics and yet have generated different ANNs. Such differences may be found in the numbers of hidden nodes, the regularization coefficients or in the weights (an effect of the stochastic training process). In addition, the genotypes may of course differ by having selected different subsets of molecular characteristics.

30

Although a cross-validation method is described in the present example, the ANNs may alternatively be trained using a Bayesian framework. Bayesian ANNs are known to

35

provide advantages in that over-fitting can be effectively avoided without having to resort to partitioning of the data set and cross-validation. An additional advantage is that confidence intervals can be assigned to the Bayesian network predictions. They also allow more straightforward comparisons to be made between ANN models of different orders.

Figure 5 illustrates suitable apparatus for performing the present invention. A computer system 200, generally indicated at 200 comprises a processor 201 by which suitable software may be executed. The processor 201 communicates with a store 202 within which the model data set is retained. Any known data storage medium may be used for this purpose to serve as read-only or read/write memory. The store 202 may also retain the model software for execution by the processor. A second temporary store 203 comprising suitable RAM devices is provided for use to store data during the execution of the modelling software. However, alternatively part of the store 202 may be used for this purpose.

Input devices such as a keyboard and mouse are generally indicated at 204 to allow the computer system to be controlled by a local operator including possible use of the modelling software.

The computer system 200 also contains a communication device 205 such as a modem for allowing remote access to the modelling software for example using the Internet. A remote user 206 may therefore request biological target characteristics predictions upon new molecules by entering suitable data describing such molecules from a remote location.

An example of the results which may be obtained using the method are shown in Table 1 below. The present method was applied to the "Selwood dataset" which produced a number of genotype models. In this particular example a "leave-one-out" cross-validation method was used.

Table 1 gives the peak model correlations (R) and cross-validated correlations (Q) and stable (i.e. easily reproducible) Q values for the top models according to the invention (denoted "INV"), as well as So and Karplus' best model for comparison ("So+Kar").

The results match those of So and Karplus (1996) under their rigid architectural constraints as can be seen from the last two lines of the table. However, when the method was allowed to range freely over different network architectures, it generated much improved results.

Table 1

Peak R	Peak Q	Stable Q	Molec. Charac	Hidden Nodes	Model
0.947	0.908	0.908	8	1	INV
0.949	0.909	0.899	7	1	INV
0.923	0.907	0.897	5	1	INV
0.952	0.902	0.893	4	2	INV
0.936	0.902	0.887	7	1	INV
0.967	0.909	0.884	5	2	INV
0.919	0.866	0.866	3	3	So+Kar
0.947	0.884	0.858	3	3	INV

CLAIMS

1. A method of predicting a target biological characteristic of a target molecule, the method comprising:
 - a) providing a data set for a number of known molecules, the data corresponding to a target biological characteristic and a number of molecular characteristics, wherein the molecular characteristics comprise at least one structural descriptor of the known molecules;
 - b) selecting a number of the molecular characteristics from the data set;
 - c) determining a relationship between the selected characteristics and the target biological characteristic;
 - d) obtaining data corresponding to the selected molecular characteristics for the target molecule; and,
 - e) predicting the biological target characteristic data for the target molecule based upon the obtained molecular characteristics data and the determined relationship.
- 2. A method according to claim 1, wherein the selected molecular characteristics include at least one of the structural descriptors.
3. A method according to claim 1 or claim 2, wherein the steps (b) and (c) are performed repeatedly.
4. A method according to any of claims 1 to 3, wherein the molecular characteristics are selected using an evolutionary algorithm.
5. A method according to any of claims 1 to 4, wherein the relationship between the target biological

characteristic and the molecular characteristics is determined using an artificial neural network.

5 6. A method according to claim 5 when dependent upon claim 4, wherein the evolutionary algorithm selects at least one artificial neural network parameter for use in accordance with the artificial neural network.

10 7. A method according to claim 6, wherein the artificial neural network has a number of hidden nodes and wherein the artificial neural network parameter is used to control the number of hidden nodes in the artificial neural network.

15 8. A method according to claim 6 or claim 7, wherein the artificial neural network has a regularisation parameter and wherein the artificial neural network parameter corresponds to the regularisation parameter.

20 9. A method according to any of the preceding claims, further comprising selecting a number of groups of molecular characteristics having corresponding determined relationships and combining the predictions of each selected group and corresponding relationship to produce a prediction for the biological target characteristic of the
25 target molecule.

30 10. A method according to any of the preceding claims, wherein the biological target characteristic is a physiological property.

11. A method according to any of the preceding claims, wherein the structural descriptor represents a substructure of at least one of the known molecules.

35 12. A computer program comprising program code means adapted to perform the method according to any of claims 1 to 11 when the computer program is run on a computer.

13. A computer program according to claim 12, embodied on a computer readable medium.

14. A system for predicting a target biological characteristic of a target molecule, the system comprising:
5 a data store for storing the data set according to any of claims 1 to 11;

a program store for storing the computer program according to claim 12; and

10 a processor for executing the computer program contained within the program store.

15. A method of predicting a biological target property of a target molecule, the method comprising the steps of:-

15 a) providing a data set for a number of known molecules, the data corresponding to a target biological characteristic and a number of molecular characteristics;

20 b) determining a relationship between a number of the molecular characteristics of the data set and the biological target property of the data set by repeatedly performing the steps of:-

i) selecting a number of groups of molecular characteristics;

25 ~~ii) selecting at least one artificial neural network parameter for each group;~~

30 iii) for each group, using an artificial neural network in accordance with the selected artificial neural network parameter to determine a relationship between the molecular characteristics data and the biological target property data of the data set;

35 iv) assessing the performance of the respective artificial neural network for each group;

v) repeatedly performing steps (i) to (iv) using the assessed performance in the

selection of molecular characteristics and artificial neural network parameters in subsequent steps (i) and (ii); and

- 5 c) predicting the biological target property for the target molecule using the determined relationship and the respective molecular characteristics of the target molecule.

10 16. A method according to claim 15, wherein the number of molecular characteristics in the data set is greater than the number of molecules in the data set.

15 17. A method according to claim 15 or claim 16, wherein the artificial neural network parameter determines the number of hidden nodes in the artificial neural network.

18. A method according to any of claims 15 to 17, wherein the artificial neural network parameter is a regularisation parameter for the artificial neural network.

20

19. A method according to any of claims 15 to 18, wherein the performance of the artificial neural networks is assessed using a fitness parameter.

~~25 20. A method according to any of claims 15 to 19, wherein the selection of the groups of molecular characteristics is performed using an evolutionary algorithm.~~

30 21. A method according to any of claims 15 to 20, wherein the selection of the artificial neural network parameter is performed using an evolutionary algorithm.

35 22. A method according to any of claims 15 to 21, wherein a number of relationships are determined between groups of selected molecular characteristics with associated artificial neural network parameters, and the biological target property, and wherein a prediction of the biological

target property for the target molecule comprises combining the predictions of the determined relationships.

23. A method according to any of claims 15 to 22, wherein the biological target characteristic is a physiological property.

24. A computer program comprising program code means adapted to perform the method according to any of claims 15 to 23 when the computer program is run on a computer.

25. A computer program according to claim 24, embodied on a computer readable medium.

26. A system for predicting a biological target property of a target molecule, the system comprising:
a data store for storing the data set according to any of claims 15 to 23;
a program store for storing the computer program according to claim 24; and
a processor for executing the computer program stored in the program store.

27. A method of determining a molecular structural descriptor from a molecular coding data set contained within a store, the method comprising:

- a) accessing the store to select first molecular coding data from the molecular coding data set, the first molecular coding data describing a first molecular structure;
- b) accessing the store to select second molecular coding data from the molecular coding data set, the second molecular coding data describing a second molecular structure;
- c) processing the selected first and second molecular codings data to determine common structural coding data representing a common

molecular structure between the first and second molecular structures; and

- d) storing the determined common structural coding data in the store, the common structural coding data representing the molecular structural descriptor.

28. A method according to claim 27, further comprising repeatedly performing steps (a) to (d) upon the coding data in the store, such that common structural coding data is determined for all of the molecular coding data and common structural coding data in the store related to the molecular coding data set.

29. A method according to claim 27 or claim 28, wherein the step of processing the first and second molecular codings data comprises the steps of:

- i) converting the first and second molecular codings data to first and second coloured graphs according to graph theory;
- ii) determining a docking graph from the first and second graphs;
- iii) identifying at least one clique within the docking graph; and
- iv) converting each clique identified into common structural coding data.

30. A method according to claim 29, wherein the common structural coding data is stored in the store only if it represents a unique molecular structure with respect to the corresponding molecular structures of the coding data already contained within the store.

31. A method according to any of claims 28 to 30, further comprising ranking the common structural coding data according to the frequency with which it is identified within the molecular coding data in the store; and

discarding the common structural coding data which occurs less frequently than a predetermined frequency threshold.

5 32. A method according to any of claims 27 to 31, wherein the determined common structural coding data is common substructural coding data.

10 33. A method according to any of claims 1 to 11, wherein the molecular structural descriptors are selected according to the method of any of claims 27 to 32.

15 34. A computer program comprising program code means adapted to perform the method according to any of claims 27 to 33, when the computer program is run on a computer.

35. A computer program according to claim 34, embodied on a computer readable medium.

20 36. A system for determining a molecular structural descriptor from molecular coding data contained within a data set, the system comprising:

a data store containing a molecular coding data set;
a program store for storing the computer program according to claim 34; and

25 ~~a processor for executing the computer program~~
contained within the program store.

30 37. A method as substantially hereinbefore described with reference to any of Figures 1 to 4 of the accompanying drawings.

38. Apparatus as substantially hereinbefore described with reference to Figure 5 of the accompanying drawings.

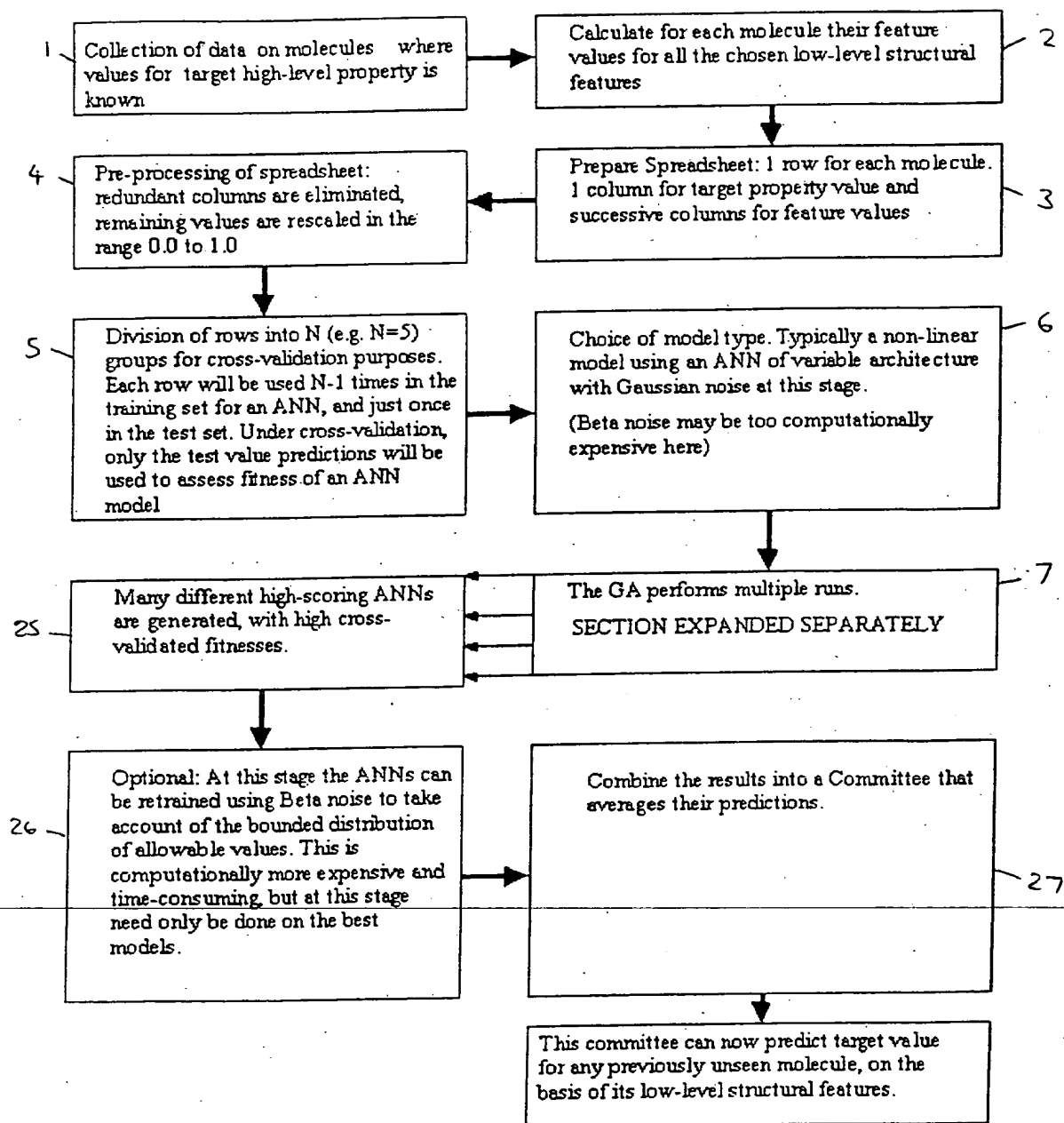


Fig 1

THIS PAGE BLANK (USPTO)

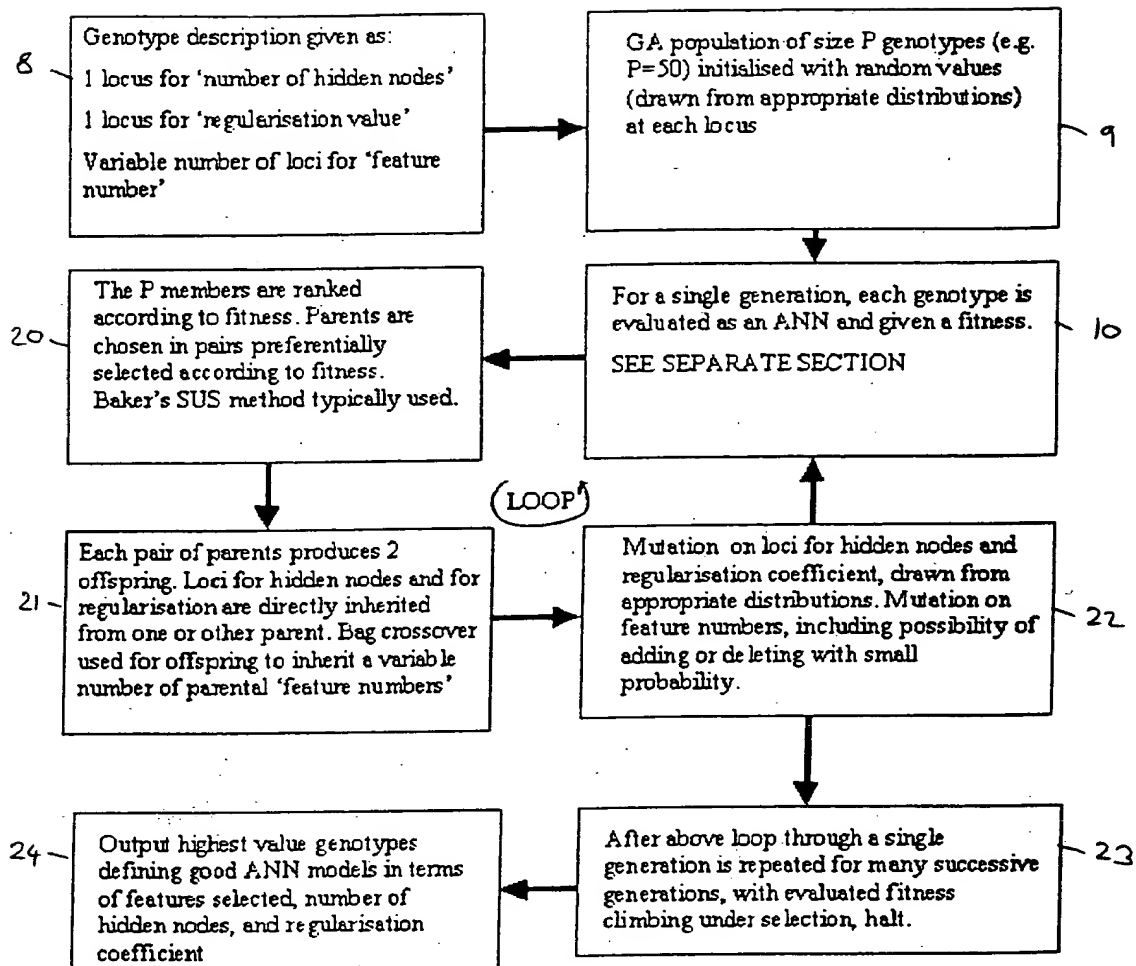


Fig. 2

THIS PAGE BLANK (USPTO)

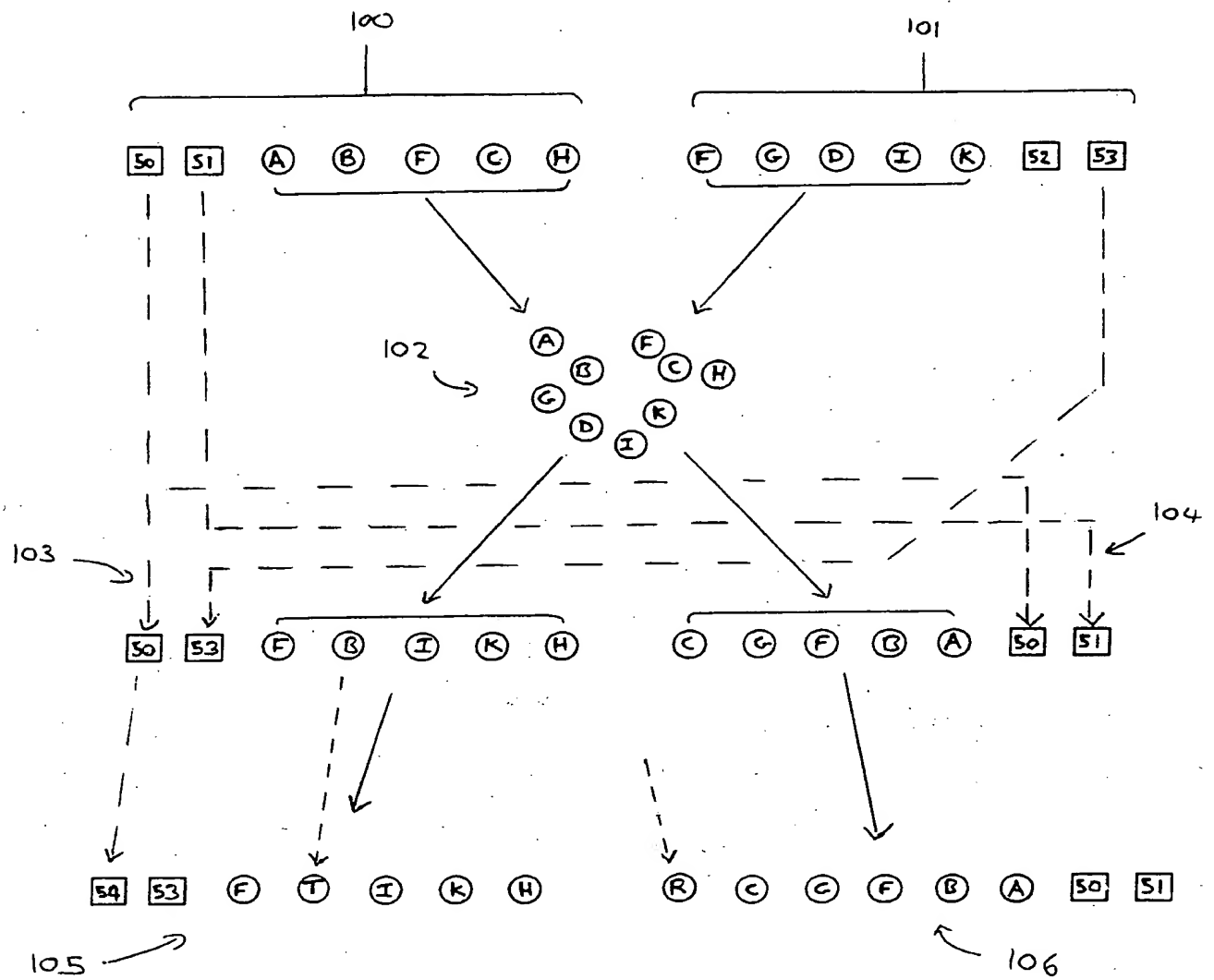


Fig. 3

THIS PAGE BLANK (USPTO)

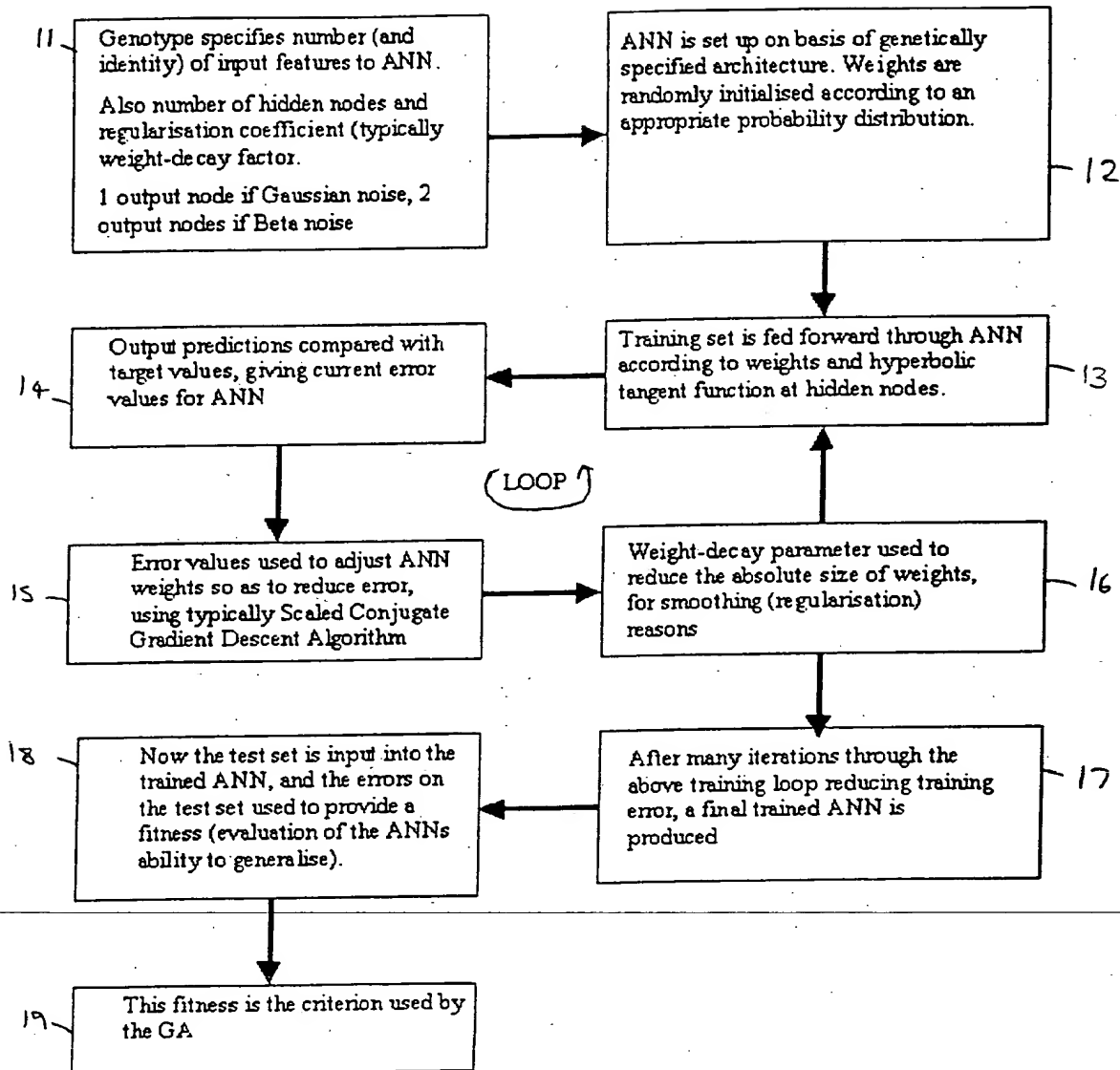


Fig. 4

THIS PAGE BLANK (USPTO)

5/5

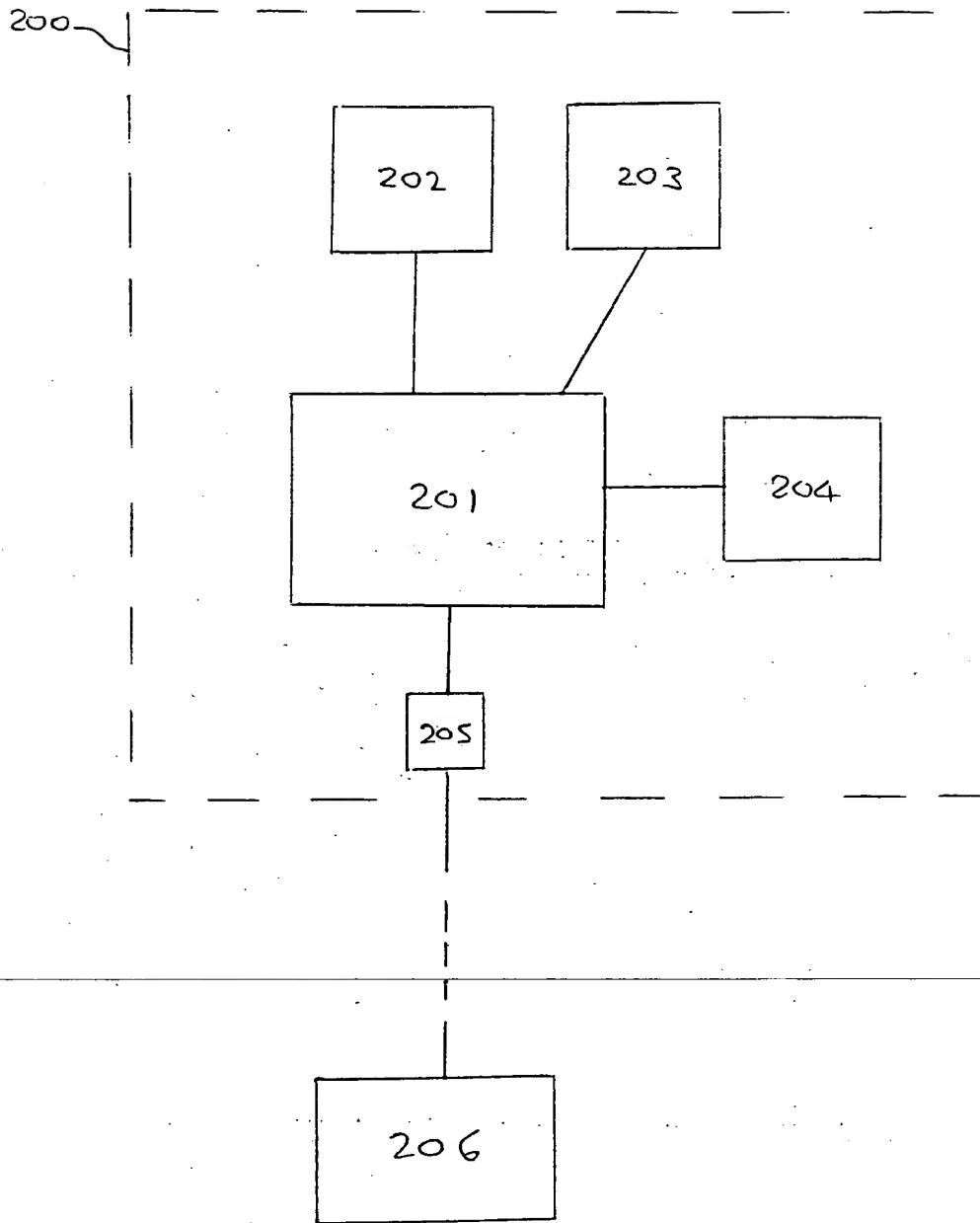


Fig. 5

THIS PAGE BLANK (USPTO)